APPENDIX P

American Educational Research Association (AERA)
POSITION STATEMENT CONCERNING
HIGH-STAKES TESTING IN PREK-12 EDUCATION
Adopted July 2000

The American Educational Research Association (AERA) is the nation's largest professional organization devoted to the scientific study of education. The AERA seeks to promote educational policies and practices that credible scientific research has shown to be beneficial, and to discourage those found to have negative effects. From time to time, the AERA issues statements setting forth its research-based position on educational issues of public concern. One such current issue is the increasing use of high-stakes tests as instruments of educational policy.

This position statement on high-stakes testing is based on the 1999 *Standards for Educational and Psychological Testing*. The *Standards* represent a professional consensus concerning sound and appropriate test use in education and psychology. They are sponsored and endorsed by the AERA together with the American Psychological Association (APA) and the National Council on Measurement in Education (NCME). This statement is intended as a guide and a caution to policy makers, testing professionals, and test users involved in high-stakes testing programs. However, the *Standards* remain the most comprehensive and authoritative statement by the AERA concerning appropriate test use and interpretation.

Many states and school districts mandate testing programs to gather data about student achievement over time and to hold schools and students accountable. Certain uses of achievement test results are termed "high stakes" if they carry serious consequences for students or for educators. Schools may be judged according to the school-wide average scores of their students. High school-wide scores may bring public praise or financial rewards; low scores may bring public embarrassment or heavy sanctions. For individual students, high scores may bring a special diploma attesting to exceptional academic accomplishment; low scores may result in students being held back in grade or denied a high school diploma.

These various high-stakes testing applications are enacted by policy makers with the intention of improving education. For example, it is hoped that setting high standards of achievement will inspire greater effort on the part of students, teachers, and educational administrators. Reporting of test results may also be beneficial in directing public attention to gross achievement disparities among schools or among student groups. However, if high-stakes testing programs are implemented in circumstances where educational resources are inadequate or where tests lack sufficient reliability and validity for their intended purposes, there is potential for serious harm.

Policy makers and the public may be misled by spurious test score increases unrelated to any fundamental educational improvement; students may be placed at increased risk of educational failure and dropping out; teachers may be blamed or punished for inequitable resources over which they have no control; and curriculum and instruction may be severely distorted if high test scores per se, rather than learning, become the overriding goal of classroom instruction.

This statement sets forth a set of conditions essential to sound implementation of high-stakes educational testing programs. It is the position of the AERA that every high-stakes achievement testing program in education should meet all of the following conditions:

Protection Against High-Stakes Decisions Based on a Single Test

Decisions that affect individual students' life chances or educational opportunities should not be made on the basis of test scores alone. Other relevant information should be taken into account to enhance the overall validity of such decisions. As a minimum assurance of fairness, when tests are used as part of making high-stakes decisions for individual students such as promotion to the next grade or high school graduation, students must be afforded multiple opportunities to pass the test. More importantly, when there is credible evidence that a test score may not adequately reflect a student's true proficiency, alternative acceptable means should be provided by which to demonstrate attainment of the tested standards.

Adequate Resources and Opportunity to Learn

When content standards and associated tests are introduced as a reform to change and thereby improve current practice, opportunities to access appropriate materials and retraining consistent with the intended changes should be provided before schools, teachers, or students are sanctioned for failing to meet the new standards. In particular, when testing is used for individual student accountability or certification,

students must have had a meaningful opportunity to learn the tested content and cognitive processes. Thus, it must be shown that the tested content has been incorporated into the curriculum, materials, and instruction students are provided before high-stakes consequences are imposed for failing examination.

Validation for Each Separate Intended Use

Tests valid for one use may be invalid for another. Each separate use of a high-stakes test, for individual certification, for school evaluation, for curricular improvement, for increasing student motivation, or for other uses requires a separate evaluation of the strengths and limitations of both the testing program and the test itself.

Full Disclosure of Likely Negative Consequences of High-Stakes Testing Programs

Where credible scientific evidence suggests that a given type of testing program is likely to have negative side effects, test developers and users should make a serious effort to explain these possible effects to policy makers.

Alignment Between the Test and the Curriculum

Both the content of the test and the cognitive processes engaged in taking the test should adequately represent the curriculum. High-stakes tests should not be limited to that portion of the relevant curriculum that is easiest to measure. When testing is for school accountability or to influence the curriculum, the test should be aligned with the curriculum as set forth in standards documents representing intended goals of instruction. Because high-stakes testing inevitably creates incentives for inappropriate methods of test preparation, multiple test forms should be used or new test forms should be introduced on a regular basis, to avoid a narrowing of the curriculum toward just the content sampled on a particular form.

Validity of Passing Scores and Achievement Levels

When testing programs use specific scores to determine "passing" or to define reporting categories like "proficient," the validity of these specific scores must be established in addition to demonstrating the representativeness of the test content. To begin with, the purpose and meaning of passing scores or achievement levels must be clearly stated. There is often confusion, for example, among minimum competency levels (traditionally required for grade-to-grade promotion), grade level (traditionally defined as a range of scores around the national average on standardized tests), and "world-class" standards (set at the top of the distribution, anywhere from the 70th to the

99th percentile). Once the purpose is clearly established, sound and appropriate procedures must be followed in setting passing scores or proficiency levels. Finally, validity evidence must be gathered and reported, consistent with the stated purpose.

Opportunities for Meaningful Remediation for Examinees Who Fail High-Stakes Tests

Examinees who fail a high-stakes test should be provided meaningful opportunities for remediation. Remediation should focus on the knowledge and skills the test is intended to address, not just the test performance itself. There should be sufficient time before retaking the test to assure that students have time to remedy any weaknesses discovered.

Appropriate Attention to Language Differences Among Examinees

If a student lacks mastery of the language in which a test is given, then that test becomes, in part, a test of language proficiency. Unless a primary purpose of a test is to evaluate language proficiency, it should not be used with students who cannot understand the instructions or the language of the test itself. If English language learners are tested in English, their performance should be interpreted in the light of their language proficiency. Special accommodations for English language learners may be necessary to obtain valid scores.

Appropriate Attention to Students with Disabilities

In testing individuals with disabilities, steps should be taken to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement.

Careful Adherence to Explicit Rules for Determining Which Students Are to be Tested

When schools, districts, or other administrative units are compared to one another or when changes in scores are tracked over time, there must be explicit policies specifying which students are to be tested and under what circumstances students may be exempted from testing. Such policies must be uniformly enforced to assure the validity of score comparisons. In addition, reporting of test score results should accurately portray the percentage of students exempted.

Sufficient Reliability for Each Intended Use

Reliability refers to the accuracy or precision of test scores. It must be shown that scores reported for individuals or for schools are sufficiently accurate to support each intended interpretation. Accuracy should be examined for the scores actually used. For example, information about the reliability of raw scores may not adequately describe the accuracy of percentiles; information about the reliability of school means may be insufficient if scores for subgroups are also used in reaching decisions about schools.

Ongoing Evaluation of Intended and Unintended Effects of High-Stakes Testing

With any high-stakes testing program, ongoing evaluation of both intended and unintended consequences is essential. In most cases, the governmental body that mandates the test should also provide resources for a continuing program of research and for dissemination of research findings concerning both the positive and the negative effects of the testing program.